IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 1, March 2024, pp. 737~746
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i1.pp737-746

Towards a new approach to maximize tax collection using machine learning algorithms

Nabil Ourdani, Mohamed Chrayah, Noura Aknin
TMS Research Unit, Abdelmalek Essaadi University, UAE, Tetuan, Morocco

Article Info

Article history:

Received Sep 25, 2023
Revised Oct 17, 2023
Accepted Nov 7, 2023

Keywords:

Clustering
Machine learning
Novel strategies
Tax debt collection
Taxpayer segmentation

ABSTRACT

Efficient tax debt collection is a challenge for Moroccan local tax authorities. This article explores the potential of machine learning techniques and novel strategies to enhance efficiency in this process. We present a practical use case demonstrating the application of machine learning for taxpayer segmentation, improving accuracy in identifying high-risk debtors. Using a comprehensive dataset of tax payment behavior, we showcase the effectiveness of machine learning algorithms in segmenting taxpayers based on their likelihood of non-compliance or debt accumulation. We also investigate innovative strategies that integrate behavioral economics principles to enable better targeted interventions. Real-world case studies in local tax debt collection highlight the impact of these strategies. The findings underscore the transformative potential of machine learning techniques and novel strategies in improving the efficiency of local tax debt collection. Accurate identification of high-risk debtors and tailored enforcement actions help maximize revenue while minimizing resource waste. This research contributes to the existing knowledge by providing insights into the implementation of machine learning techniques and novel strategies in tax debt collection. It emphasizes the importance of data-driven approaches and highlights how local tax authorities can drive efficiency and optimize revenue collection by embracing these advancements.

Corresponding Author:

Nabil Ourdani
TMS Research Unit, Abdelmalek Essaadi University
Tetuan, Morocco
Email: nabil.ourdani@etu.uae.ac.ma

1. INTRODUCTION

In recent years, local tax authorities in Morocco have faced significant challenges in collecting tax debts, resulting in substantial revenue loss and hindering the provision of essential public services. According to a report by the Court of Accounts of Morocco [1], there was a 29% deterioration in revenue collection as outstanding amounts to be recovered increased from 13 billion Moroccan Dirhams (MAD) to 16.8 billion MAD between 2009 and 2013. This increase, with an average annual growth rate of 7.3%, indicates significant challenges in the collection process. Tax debt collection involves complex processes, including taxpayer segmentation, enforcement actions, and resource allocation (human, financial, and technological), particularly tailored to the unique context of African countries [2]. However, traditional approaches often suffer from inefficiencies, relying on manual methods and generic strategies that do not effectively target high-risk debtors. To address these challenges and achieve a reduction in processing time, an increase in taxes collected, and a decrease in the number of legal disputes, there is a growing interest in harnessing machine learning techniques and innovative strategies aimed at optimizing efficiency in local tax debt collection [3].

Machine learning has revolutionized various domains by enabling the automatic extraction of insights and patterns from large datasets especially in finance [4]. In the context of tax debt collection, machine learning techniques offer immense potential for improving segmentation accuracy, predicting debt default probabilities, and optimizing resource allocation [5]. By utilizing historical data and sophisticated algorithms, local tax authorities can identify taxpayers with the highest likelihood of non-compliance or debt accumulation, enabling targeted interventions and enforcement actions.

Moreover, emerging novel strategies go beyond conventional approaches to tax debt collection. These strategies incorporate behavioral economics principles [6], social network analysis, and big data analytics to gain a deeper understanding of taxpayer behavior and compliance patterns. By incorporating these novel strategies, local tax authorities can predict taxpayer decisions [7] and develop proactive and tailored approaches that address the root causes of non-compliance and promote voluntary tax payment.

Tax collection is of utmost importance to governments as it provides essential funding for public services, making the continuous improvement of tax collection methods and strategies crucial. Several authors have explored the role of behavioral economics in designing and enhancing tax policy and administration [8]. Walsh [9] explores the potential for tax services to benefit from the behavioral study of taxpayers and its implications for policy-making and budgeting. The research delves into the opportunities presented by understanding taxpayer behavior and how it can inform the development of more effective tax policies.

Another relevant work by Congdon *et al*. [10] delve into the application of behavioral economics within the context of tax policy transformation. They examine how insights from behavioral economics can inform the implementation of effective tax policies that consider individuals' behavioral biases and decision-making processes. By incorporating principles from behavioral economics, policymakers can design interventions that encourage voluntary compliance, minimize tax evasion, and optimize revenue collection. Kling and Mullainathan emphasize the importance of understanding the behavior of taxpayer.

Other studies have also examined taxpayer segmentation as a strategy to enhance tax debt collection. A study by Stankevicius *et al*. [11] presents a conceptual model for taxpayer segmentation based on behavior and legal information. The model aims to enhance tax administration and compliance strategies by identifying distinct taxpayer segments. It considers factors such as payment history and compliance patterns to classify taxpayers. The model enables targeted interventions, resource allocation, and personalized tax policies. It provides insights into taxpayer characteristics and behaviors. The study highlights the importance of segmentation for effective tax administration. The model offers a framework to analyze and classify taxpayers efficiently. By implementing the model, tax authorities can improve compliance and revenue collection. The research contributes to more efficient tax policies and administration.

Related to segmenting taxpayers, recent studies like Tymchenko *et al*. [12], and Fox *et al*. [13], suggests that there are a variety of factors that can be used to segment taxpayers into different groups, including risk of noncompliance, compliance motivation, and perception of fairness. Other recent studies, such the one conducted by Vasco *et al*. [14] used data mining techniques to study the Spanish Personal Income Tax sample designed by the Institute for Fiscal Studies.

These works collectively demonstrate the potential of machine learning in taxation and its integration with economic principles to improve fraud detection, tax evasion prediction, tax debt collection, taxpayer segmentation and make novel strategies. As a continuation, and in the hope of covering the lack for the field of local finances, the objective of this article is to explore the potential of machine learning techniques in driving efficiency in local tax debt collection and analyze the associated benefits and practical implications. We will examine different machine learning techniques for taxpayer segmentation and novel strategies in local tax debt collection to provide valuable insights and lessons learned.

By delving into the new frontiers of machine learning and novel strategies, local tax authorities can enhance their capacity to effectively identify high-risk debtors, optimize resource allocation, and improve the overall efficiency of tax debt collection processes. This article aims to contribute to the existing body of knowledge by shedding light on the transformative potential of these approaches in the context of local tax debt collection. Through this article, we aim to inspire further research, collaboration, and implementation of these innovative approaches to overcome the challenges faced by local authorities in tax debt collection. By harnessing the power of data-driven insights, local tax authorities can improve taxpayer segmentation, enhance enforcement actions, and ultimately maximize revenue collection.

2. COMPREHENSIVE THEORETICAL BASIS

The proposed work is built upon a solid foundation of theoretical bases that underpin its framework. In this section, we explore the key theoretical concepts and frameworks that inform our research. Firstly, we we draw upon taxpayer segmentation concept, reasons of use and techniques available. Secondly, we delve into the field of behavioral economics, which provides insights into how individuals make economic decisions

and respond to incentives. By integrating principles from behavioral economics, we aim to understand the underlying factors that influence taxpayer behavior and compliance.

2.1. Taxpayer segmentation

Taxpayer segmentation is a process employed by governments and tax authorities to categorize taxpayers based on specific characteristics or traits. This segmentation approach acknowledges the fact that taxpayers are not uniform and that different groups may display diverse compliance behaviors, risks, or needs. Segmentation can be based on various factors, including the type and size of the business, the industry in which it operates, the geographical location, or the taxpayer's compliance track record [9].

Some examples of taxpayer segmentation strategies that have been used by governments and tax authorities include risk-based audits, targeted outreach and education programs, and specialized support for small businesses [15]. Most reasons of using taxpayer segmentation [16] are:

– Improved efficiency: By segmenting taxpayers into different groups, tax authorities can more effectively target their resources and efforts towards specific groups of taxpayers. This can help to reduce the burden on taxpayers and improve overall efficiency.
– Enhanced compliance: By identifying and targeting specific groups of taxpayers that may be more likely to engage in non-compliant behavior, tax authorities can improve compliance rates and increase tax revenues.
– Customized support: By segmenting taxpayers into different groups, tax authorities can provide customized support and assistance to taxpayers who may need it. For example, they may provide more resources or guidance to small businesses or first-time taxpayers.

Several techniques can be employed for taxpayer segmentation, including:

– Data analysis and modeling: Tax authorities can use data analysis and modeling techniques to identify patterns and trends among taxpayers, which can be used to group taxpayers into different segments.
– Risk assessment: Tax authorities can use risk assessment techniques to identify taxpayers who may be more likely to engage in non-compliant behavior or to underestimate their tax liabilities. These taxpayers can be targeted for additional scrutiny or outreach.
– Customer segmentation: Tax authorities can use customer segmentation techniques to group taxpayers based on factors such as their needs, preferences, and behaviors. This can help to tailor outreach and assistance programs to specific groups of taxpayers.
– Cluster analysis: Cluster analysis is a statistical technique that can be used to group similar taxpayers together based on shared characteristics.
– Decision trees: Decision tree analysis is a technique that can be used to segment taxpayers based on a series of decisions or rules.

This process can inform targeted tax enforcement strategies and improve the efficiency and effectiveness of tax collection. Previous research on taxpayer segmentation [11] has focused on a variety of factors, such as demographic characteristics, tax compliance behaviors, and economic indicators, and has used various methods, including statistical analysis and machine learning.

2.2. Novel strategies

Taxpayer segmentation refers to the process of categorizing taxpayers into distinct groups based on specific characteristics or behaviors. The goal of segmentation is to identify high-risk debtors who are more likely to default on their tax obligations or accumulate tax debt. Machine learning techniques revolutionize taxpayer segmentation by leveraging advanced algorithms to analyze vast amounts of data and extract patterns and insights that may not be apparent through manual methods. Machine learning algorithms can identify complex relationships, nonlinear patterns, and interactions among various variables, enabling more accurate and granular taxpayer segmentation. These techniques improve segmentation accuracy by considering a wide range of features, including financial indicators, historical tax payment behavior, demographic information, and other relevant data points.

However, machine learning alone is not sufficient to drive optimal tax debt collection. This is where novel strategies come into play. Novel strategies encompass innovative approaches that go beyond traditional methods and incorporate additional elements to enhance taxpayer compliance and debt collection outcomes. These strategies often draw upon insights from fields such as behavioral economics, social network analysis, and big data analytics. Behavioral economics principles consider the psychological and behavioral factors that influence taxpayer decision-making [6]. By understanding cognitive biases, social norms, and psychological triggers, tax authorities can design interventions and communication strategies that influence taxpayer behavior, promote voluntary compliance, and reduce tax debt accumulation.

Social network analysis examines the relationships and connections among taxpayers within a network. By analyzing social connections, business relationships, or transactional networks, revenue agencies

can identify clusters of high-risk debtors or uncover hidden non-compliance patterns [17]. This information enables targeted enforcement actions and facilitates the identification of tax evasion schemes or collusion.

Big data analytics leverages the processing power and analytical capabilities to handle large and diverse datasets. By integrating and analyzing multiple data sources, including financial records, transactional data, social media, or external databases, revenue agencies can gain deeper insights into taxpayer behavior, detect trends, anomalies, or correlations that inform targeted interventions and resource allocation for tax debt collection [18]. One of those strategies is nudging [19], [20].

These novel strategies offer innovative approaches to tax debt collection by considering behavioral factors, network relationships, comprehensive data analysis, and facilitating more efficient personalized communication [21]. Taxpayer segmentation forms the foundation for efficient tax debt collection. Machine learning techniques enhance segmentation accuracy by analyzing complex patterns and variables. Novel strategies further enhance the effectiveness of taxpayer segmentation and contribute to the development of tailored interventions and resource allocation strategies, ultimately driving improved tax debt collection outcomes [22].

Existing literature and research demonstrate the effectiveness and potential of novel strategies based on machine learning techniques in enhancing taxation policy [23], [24] and compliance [19]. By leveraging personalized interventions, incorporating behavioral economics principles, and optimizing resource allocation, local tax authorities can achieve higher compliance rates, improve collection outcomes, and ensure the provision of essential public services. These studies provide valuable insights for policymakers seeking to adopt innovative approaches to tax debt collection.

3. METHODOLOGY

The primary objective of our proposed methodology is to establish the groundwork for an effective advisory system designed to offer reminders and facilitate prosecutions. This system is intended to be a valuable resource for Moroccan local tax collectors, enabling them to enhance their performance in recovering territorial claims efficiently. By implementing this system, we aspire to contribute significantly to the overall improvement of tax collection processes in Morocco.

3.1. Proposed framework approach

A proposal has been made to enhance dunning and collection policies through the implementation of a novel machine learning analysis approach. This approach aims to improve debt recovery and collection outcomes and involves several key steps:

- Data collection and integration: Collect pertinent information from diverse integrated sources including taxpayer records, payment history, communication logs, and demographic details, and consolidate it into a comprehensive dataset suitable for analysis.
- Feature engineering: Identify and create meaningful features from the collected data that can provide insights into taxpayer behavior and payment patterns. This may include variables like payment frequency, average payment amount, time since last payment, taxpayer score, and any other relevant factors.
- Model development: Utilize machine learning algorithms to develop models that can analyze the collected data and identify patterns related to payment behavior.
- Segmentation and risk profiling: Apply clustering algorithms to segment taxpayer into different groups based on their payment’s behaviors. This segmentation can help in targeting appropriate collection strategies and resources for each taxpayer group.
- Predictive analytics: Use the developed models to predict future payment behavior for individual taxpayers. This can help in identifying taxpayers who are likely to default on their payments or accumulate more debt. By predicting taxpayer behavior, collection efforts can be prioritized and tailored accordingly.
- Intervention design: Utilize insights from behavioral economics and other relevant fields to design targeted interventions and communication strategies. These interventions can leverage psychological triggers, incentives, and personalized messaging to influence taxpayer behavior and promote timely payments.
- Optimization and monitoring: Continuously monitor and evaluate the performance of the machine learning models and collection strategies. Regularly update and refine the models based on new data and feedback to improve accuracy and effectiveness.
- Collaboration and knowledge sharing: Foster collaboration between data analysts, collection teams, and other stakeholders to share insights and best practices. This collaboration can help in implementing the findings from machine learning analysis effectively and aligning collection policies with business goals.

By implementing these steps, local and national tax authorities can leverage machine learning and novel strategies to enhance their dunning and collection policies. The use of predictive analytics, taxpayer

segmentation, and personalized interventions can lead to improved debt recovery rates and optimized resource allocation for collection efforts.

3.2. Practical variation overview

The technical implementation of the proposed solution to enhance dunning and collection policies through a novel machine learning analysis approach involves several key steps while considering the constraints of data availability at the local authority level. Initially, a comprehensive experimental dataset will be collected, integrating relevant taxpayer payment history. Subsequently, feature engineering techniques will be applied to extract meaningful variables that provide insights into taxpayer behavior and payment patterns. These variables will be used to generate a score for each taxpayer. Multiple machine learning algorithms will then be compared to determine the segmentation of taxpayers based on their profiles. This segmentation will enable the implementation of targeted collection strategies and the development of a recommendation system core that optimizes recovery efforts.

The performance of the implemented machine learning models and collection strategies should be continuously monitored and optimized based on new data and feedback. This iterative process ensures a continuous improvement in debt recovery and collection outcomes. Figure 1 illustrates a high-level overview of the technical pipeline adopted. Detailed explanations of the mentioned algorithms will be presented in a later section.

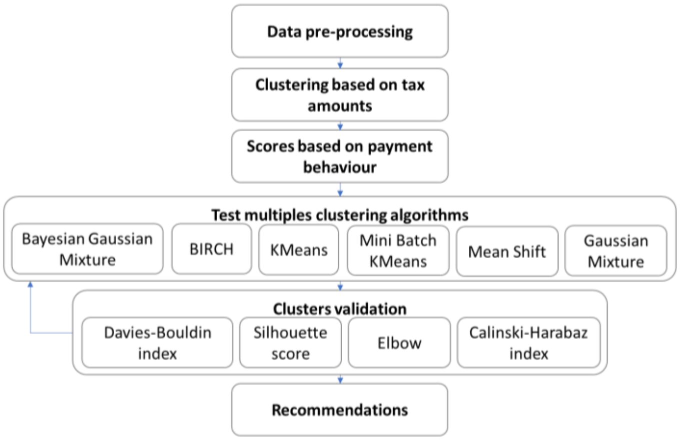

Figure 1. High level pipeline

3.3. Experimental data set

The application of the proposed methodology relies on an anonymized dataset obtained from a specific local authority in Morocco. The dataset focuses specifically on records pertaining to the municipal services tax. It encompasses a multitude of individual tax records at the local level, spanning from 2010 to 2021, and contains a total of 2,739,934 entries. These records include technical identifiers of taxpayers, as well as information regarding the type of taxpayer (moral or physical person). Additionally, the dataset provides relevant details about the taxpayers' debts, amount and dates including information on payment statuses. Figure 2 provides further details regarding the dataset. At the data preprocessing stage, we clean and preprocess the data to remove missing values, outliers, and irrelevant features.

3.4. Clustering and validation methods

The application of our proposed methodology involves comparing various clustering models utilizing machine learning algorithms. Clustering algorithms serve as fundamental tools in data analysis and machine learning, facilitating the grouping of data points based on their inherent similarities. To assess their performance, we employ validation metrics, which aid in the selection of the most suitable clustering model.

In our study, we employ prominent clustering algorithms, including the Gaussian mixture model (GMM) [25], which models data using a mixture of Gaussian distributions, and the variational Bayesian Gaussian mixture model (VBGMM) [26], renowned for its probabilistic clustering approach. Additionally, we utilize balanced iterative reducing and clustering using hierarchies (BIRCH) for hierarchical clustering [27]. The Kmeans model [28], mini batch Kmeans model [29], and mean shift model [30] offer versatile and effective methods for partitioning data into clusters. These algorithms play pivotal roles in various applications, such as taxpayer segmentation, anomaly detection, and image processing, by revealing hidden patterns within data.

To assess the quality of our taxpayer segmentation solution, we utilize four pertinent metrics: Davies-Bouldin Index [31], Silhouette Score [32], Elbow method [33]**Error! Reference source not found.** and C alinski-Harabasz Index [34]. These metrics provide valuable insights into the performance of our clustering models. Also, they guide our selection of the most appropriate approach for taxpayer segmentation and other data-driven tasks.

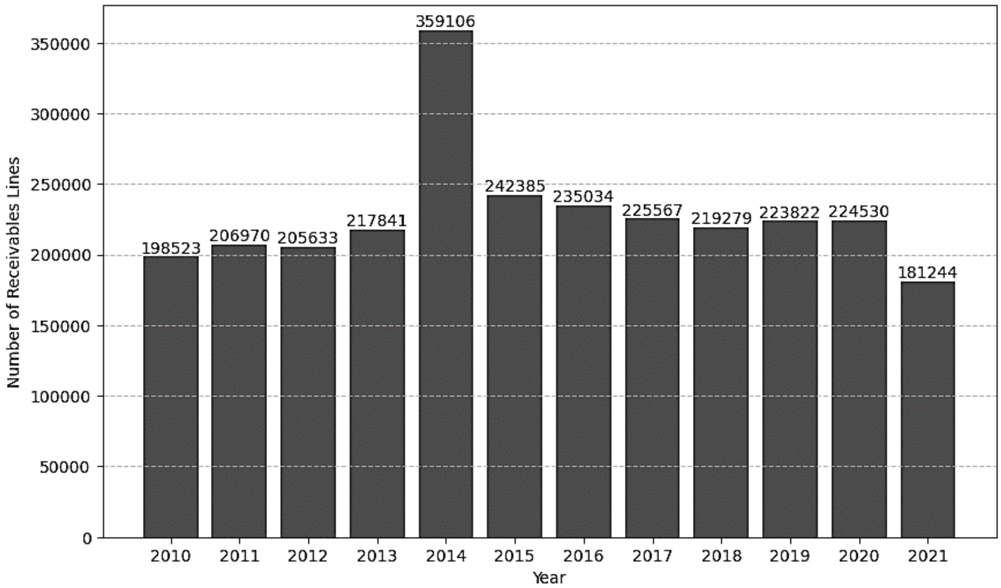

Figure 2. Number of receivables lines by years

4. APPLICATION AND RESULTS

The first segmentation criterion employed, utilizing existing information within the dataset, focuses on the "amount" of issued receivables. This approach enabled us to enhance our dataset by introducing an "amount class" column. To differentiate the potential clusters formed based on the amount parameter, we conducted comparisons by applying various algorithms as mentioned and performed model comparisons using metrics.

Based on the Silhouette criterion, the most suitable model is mini batch Kmeans, while the Calinski-Harabasz index indicates that Kmeans performs best. Regarding the Davies Bouldin index, BIRCH emerges as the top model. Notably, Kmeans appears to be the most favorable choice across all three criteria. As a result of this analysis, taxpayers have been categorized into four clusters: small, medium, large, and very large. The outcomes of this segmentation are detailed in the Table 1.

Table 1. Taxpayer segmentation according to “amount”

Debt amount	Class	Count	Cluster
63.9 <= Amount <= 30,870	Small	65,325	Cluster 1
30,968.2 <= Amount <= 192,570	Medium	758	Cluster 0
201,600 <= Amount <= 600,915	Big	51	Cluster 2
710,181.88 <= Amount <= 1,400,690.6	Very big	5	Cluster 3

In our study, receivables are marked by a life cycle at the collection agent level which begins with the management of the claim when it is created (annotated by ‘PEC’) and goes through several statuses depending on the events concerning it. These events can be payments, reminders by free notices (annotated by ‘DASF’), orders, non-value proposals (annotated by ‘PNV’) or non-value admissions (annotated by ‘ANV’). In this paper, those states are encoding by 4-bit integers. For the other state possibilities, we have set them to the value: -100. Table 2 explain the specific score given for each state. In order to have a meaningful and logical score, we take the life cycle and at each stage the score is updated depending on whether the receivable is paid (we increase the score) or not (we decrease the score). Figure 3 illustrates the scoring technique used in the proposal, indicating points added or subtracted from the score.

Table 2. State-score correspondence

Binary Code	Debt status	Score	Binary Code	Debt status	Score
1111	paid off before the lawsuits	1	1001	‘ANV’ after ‘DASF’	-8

0011	Supported ('PEC')	0	1101	Paid after Order	-9
1100	paid off after 'PNV' or 'ANV'	-3	0001	Ordered	-10
1110	paid off after 'DASF'	-4	1010	'PNV' after Order	-11
0010	'DASF' sended	-5	1000	'ANV' after Order	-13
1011	'PNV' after 'DASF'	-6			

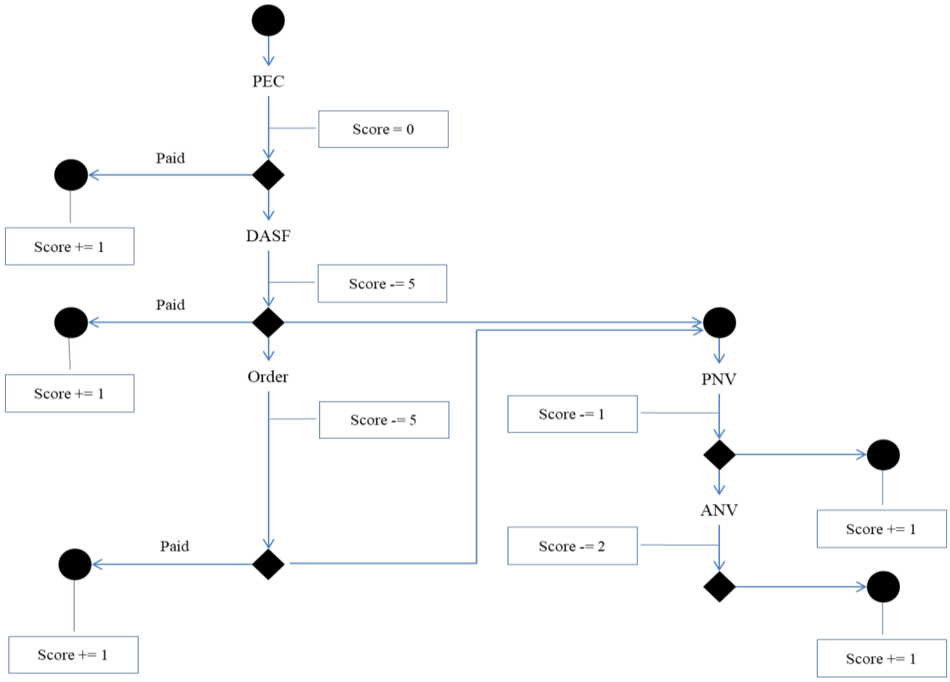

Figure 3. Proposal scoring technique

We calculate scores for all taxpayers and use them for segmentation. To distinguish potential clusters based on these scores, we compared various algorithms. The Silhouette criterion and the Calinski-Harabasz index identified Kmeans and mini batch Kmeans as the best models in this case. Meanwhile, the Davies Bouldin index favored BIRCH as the best model. However, among all the models, Kmeans appears to be the most suitable. In Table 3, taxpayers are classified into three clusters: 'bad,' 'good,' and 'excellent,' representing the final outcome.

Table 3. Taxpayer segmentation according to "score"

Score	Class	Count	Cluster
-10 <= Score <= -6.08	Bad	8,557	Cluster 1
-6 <= Score <= -1.4	Good	15,786	Cluster 2
-1.36 <= Score <= 1	Excellent	41,796	Cluster 0

After calculating the score for each taxpayer based on their payment behavior, we create a new dataset that includes the taxpayer IDs, scores, score classes, the average tax amounts and a score about this amount. Also, based on the status of each tax indicated in the initial dataset lines, we were able to enrich our dataset by the total unpaid taxes for each taxpayer (annotated by RAR). In the same way, and in order to distinguish the different clusters likely to be formed, we carried out the comparison by applying the different algorithms cited. The Figure 4 illustrates the result of the comparison performed.

According to the Silhouette criterion, the best model is mini batch Kmeans. For Calinski-Harabasz, the best is Kmeans. For Davies Bouldin, the best role model is BIRCH. Among all models Mini batch Kmeans seems to be the best according to three criteria. The obtained results reveal distinct clusters based on 'Amount' and also in relation to the outstanding amounts to be recovered. This provides decision-makers with an opportunity, for instance, to target taxpayers with a favorable behavior score who still have outstanding amounts to pay. Table 4 summarize the obtained results.

Model	Sil_score	CH_score	DV_score	setting
mini batch kmeans	0.791162	316781.207296	0.336086	{'n_clusters': 3}
BGMM	0.643834	151475.844468	1.106907	{'n_components': 4, 'covariance_type': 'full',...
BIRCH	0.780830	278348.152760	0.325238	{'n_clusters': 4, 'threshold': 0.001}
Kmeans	0.791155	316784.438639	0.336636	{'n_clusters': 3, 'init': 'k-means++', 'n_init...
Mean Shift	0.734756	251553.214385	0.411599	{'n_clusters': 5}
GMM	0.634652	137829.064202	1.104001	{'n_components': 4, 'covariance_type': 'full',...

Figure 4. Final models comparison

Table 4. Summary of the results obtained

Cluster name			%
Cluster 1 (41,733)	Amount	[63.9 : 1,400,690.60]	
	Amount Class	Small = 41,172	98.65 %
		Medium = 520	1.24 %
		Big = 36	0.08 %
		Very Big = 5	0.011%
	Score	[-1.36 : 1]	
	Score Class	Bad = 0	0 %
		Good = 0	0 %
		Excellent = 41,733	100 %
	Amount of debts to be recovered	[0 : 1,604,410.48]	
Cluster 3 (8,533)	Amount	[105.1 : 441,000.0]	
	Amount Class	Small = 8,444	98.95 %
		Medium = 83	0.97 %
		Big = 6	0.07 %
		Very Big = 0	0 %
	Score	[-1.36 : 1]	
	Score Class	Bad = 8,518	99.82 %
		Good = 10	0.117 %
		Excellent = 5	0.05 %
	Amount of debts to be recovered	[0 : 1,604,410.48]	
Cluster 2 (15,735)	Amount	[100.2: 356,372.1]	
	Amount Class	Small = 15,575	98.98 %
		Medium =151	0.95 %
		Big = 9	0.057 %
		Very Big =0	0%
	Score	[-6: -1.40]	
	Score Class	Bad = 0	0%
		Good = 15,735	100%
		Excellent=0	0%
	Amount of debts to be recovered	[0 : 1,381,269.71]	

Based on the analysis results, the clustering process yielded three distinct clusters. 'Cluster 1' comprised 41,733 occurrences and consisted of taxpayers from various categories (small, medium, large, very large), with a significant majority (98.6%) being small taxpayers. These small taxpayers exhibited scores ranging from -1 to 1, and impressively, all of them had excellent scores.

The second cluster, denoted as 'Cluster 2', contained 15,735 occurrences and exclusively included small taxpayers. Their scores ranged from -6 to -1.40, but remarkably, all of them had good scores, indicating a positive payment behavior. Lastly, 'Cluster 3' encompassed taxpayers belonging to the small category, with a total number of occurrences not specified. These taxpayers exhibited a poor score range of -10 to 1, with approximately 99.82% in the bad score category. Based on these findings, a recommended approach is to initially focus on 'Cluster 1' by initiating simple callback messages. Additionally, it is suggested to pay closer attention to taxpayers who belong to the "Tall and very tall" category and have scores between -1 and 1, which consists of 41 taxpayers.

The second recommendation involves addressing 'Cluster 2' and subsequently addressing 'Cluster 3' at a later stage, considering their respective characteristics and scores. These recommendations aim to optimize collection strategies and enhance debt recovery outcomes based on the observed clustering patterns and score distributions. The proposed novel strategy is rooted in nudging and suggests sending SMS messages tailored to different taxpayer segments to engage them based on their behavior and payment history.

5. CONCLUSION

The collection of territorial taxes is intricate, demanding a profound comprehension of taxpayer behavior to maintain compliance and devise effective strategies for re-engaging non-compliant taxpayers. Machine learning offers a valuable approach to scrutinize taxpayer behavior, particularly in optimizing out-of-court payment processing, where targeted interventions can significantly enhance collection outcomes. Our study on local tax debt collection achieved remarkable results, highlighting the potential benefits of innovative strategies and machine learning. We implemented a comprehensive taxpayer engagement program, involving the delivery of customized and concise SMS messages based on our study's findings. This approach yielded an impressive 84% payment rate, underscoring the efficacy of personalized communication. Taxpayers expressed feeling heard and understood, leading to heightened compliance and reduced disputes. Concurrently, the tax authority cultivated stronger relationships with taxpayers, fostering trust and transparency. Machine learning algorithms hold substantial promise in augmenting tax debt collection through personalized nudging strategies. By analyzing taxpayer behavior, tailoring messages, and predicting compliance patterns, these algorithms can optimize collection efforts, boost voluntary tax payment, and ensure efficient resource allocation. Embracing these technological advancements and innovative approaches empowers local tax authorities to maximize revenue collection and enhance the delivery of essential public services with greater efficiency. The proven success at the local level serves as a persuasive case study for regional and national tax authorities, urging them to consider the integration of machine learning and data analytics into their debt collection processes. Additionally, it promotes further research and collaboration in this field to refine and adapt these approaches to meet the unique needs and complexities of larger administrative structures.

ACKNOWLEDGEMENTS

M. Ourdani, the author, expresses deep gratitude and appreciation for the invaluable support and guidance provided by her thesis supervisors, the ITMS Research Unit members, and the esteemed professors at the University Abdelmalek Essaadi in Tetuan, Morocco. It is important to highlight that this project was undertaken within the context of a doctoral thesis and did not receive any external research funding or grants.

REFERENCES

[1] "Report on the assessment of local taxation (in french) [Rapport sur l'évaluation de la fiscalité locale]," *Court of Audit*, 2023, [Online]. Available: https://www.courdescomptes.ma/publication/rapport-sur-levaluation-de-la-fiscalite-locale/
[2] O. Okunogbe and F. Santoro, "Increasing Tax Collection in African Countries: The Role of Information Technology," *Journal of African Economies*, vol. 32, no. Supplement_1, pp. I57–I83, Mar. 2023, doi: 10.1093/jae/ejac036.
[3] J. Ordóñez and M. Hallo, "Detecting Atypical Behaviors of Taxpayers with Risk of NonPayment in Tax Administration, A Data Mining Framework," *Revista Politecnica*, vol. 52, no. 1, pp. 35–44, 2023, doi: 10.33333/rp.vol52n1.04.
[4] D. Broby, "The use of predictive analytics in finance," *The Journal of Finance and Data Science*, vol. 8, pp. 145–161, Nov. 2022, doi: 10.1016/j.jfds.2022.05.003.
[5] M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan, and P. Hajek, "Tax Default Prediction Using Feature Transformation-Based Machine Learning," *IEEE Access*, vol. 9, pp. 19864–19881, 2021, doi: 10.1109/ACCESS.2020.3048018.
[6] N. Barberis, "Richard Thaler and the Rise of Behavioral Economics," *The Scandinavian Journal of Economics*, vol. 120, no. 3, pp. 661–684, Jul. 2018, doi: 10.1111/sjoe.12313.
[7] O. Plonsky *et al.*, "Predicting human decisions with behavioral theories and machine learning," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06866
[8] B. Torgler, "Center for Research in Economics , Management and the Arts IT IS ABOUT BELIEVING :," *Zürich, CREMA Working Paper*, 2021, [Online]. Available: http://hdl.handle.net/10419/246008
[9] K. Walsh, "Understanding taxpayer behaviour - new opportunities for tax administration," *Economic and Social Review*, vol. 43, no. 3, pp. 451–475, 2012, doi: Understanding Taxpayer Behaviour – New Opportunities for Tax Administration.
[10] W. J. Congdon, J. R. Kling, and S. Mullainathan, "Behavioral Economics and Tax Policy," *National Tax Journal*, vol. 62, no. 3, pp. 375–386, Sep. 2009, doi: 10.17310/ntj.2009.3.01.
[11] E. Stankevicius and K. Kundeliene, "Theoretical Approach to Taxpayers' Segmentation," May 2017. doi: 10.3846/cbme.2017.067.
[12] O. Tymchenko, Y. Sybirianska, and A. Abramova, "The approach to tax debtors segmentation," *Ikonomicheski Izsledvania*, vol. 28, no. 5, pp. 103–119, 2019.
[13] W. F. Fox, E. P. Hargaden, and L. A. Luna, "Statutory incidence and sales tax compliance: Evidence from Wayfair," *Journal of Public Economics*, vol. 213, 2022, doi: 10.1016/j.jpubeco.2022.104716.
[14] M. D. C. G. Vasco, M. J. D. Rodríguez, and S. D. L. Santos, "Segmentation of Potential Fraud Taxpayers and Characterization in Personal Income Tax Using Data Mining Techniques," *Hacienda Publica Espanola*, vol. 239, no. 4, pp. 127–157, 2021, doi: 10.7866/HPE-RPE.21.4.4.
[15] A. Chooi, "Mobilizing Revenue: Emerging Approaches to Managing and Collecting Tax Debt to Improve Tax Payment Compliance," *The Governance Brief*, no. 49, 2023, doi: 10.22617/BRF230068.
[16] W. J. Crandall, A. Masters, and E. Gavin, "ISORA 2018: Understanding Revenue Administration," *Departmental Papers*, vol. 2021, no. 025, p. 1, 2021, doi: 10.5089/9781513592930.087.
[17] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, Apr. 2004, doi: 10.1109/MC.2004.1297301.
[18] A. Veit, "Swimming upstream: Leveraging data and analytics for taxpayer engagement - An Australian and international perspective," *eJournal of Tax Research*, vol. 16, no. 3, pp. 474–499, 2019.
[19] J. Alm, L. Burgstaller, A. Domi, A. März, and M. Kasper, "Nudges, Boosts, and Sludge: Using New Behavioral Approaches to Improve Tax Compliance," *Economies*, vol. 11, no. 9, p. 223, Sep. 2023, doi: 10.3390/economies11090223.

[20] J. A. List, M. Rodemeier, S. Roy, and G. Sun, "Judging Nudging: Understanding the Welfare Effects of Nudges Versus Taxes," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4448251.

[21] M. Ghaffari, M. Kaniewicz, and S. Stricker, "Personalized Communication Strategies: Towards a New Debtor Typology Framework," *Psychology and Behavioral Sciences*, vol. 10, no. 6, p. 256, 2021, doi: 10.11648/j.pbs.20211006.20.

[22] A. Antinyan and Z. Asatryan, "Nudging for Tax Compliance: A Meta-Analysis," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3680357.

[23] P. Battiston, S. Gamba, and A. Santoro, "Optimizing Tax Administration Policies with Machine Learning," University of Milano-Bicocca, Department of Economics, 2020. [Online]. Available: https://ideas.repec.org/p/mib/wpaper/436.html

[24] M. Andini, E. Ciani, G. de Blasio, A. D'Ignazio, and V. Salvestrini, "Targeting with machine learning: An application to a tax rebate program in Italy," *Journal of Economic Behavior & Organization*, vol. 156, pp. 86–102, Dec. 2018, doi: 10.1016/j.jebo.2018.09.010.

[25] C. M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)," 2006.

[26] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, Mar. 2006, doi: 10.1214/06-BA104.

[27] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH," *ACM SIGMOD Record*, vol. 25, no. 2, pp. 103–114, Jun. 1996, doi: 10.1145/235968.233324.

[28] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[29] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, Apr. 2010, pp. 1177–1178. doi: 10.1145/1772690.1772862.

[30] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002, doi: 10.1109/34.1000236.

[31] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[33] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267–276, Dec. 1953, doi: 10.1007/BF02289263.

[34] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics - Simulation and Computation*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610917408548446.

BIOGRAPHIES OF AUTHORS

Nabil Ourdani is an IT engineer and PhD student at Abdelmalek Essaadi University in Tetuan, Morocco. He is an experienced professional with a background in computer science, public management, and extensive training in various technologies, including cloud computing, blockchain, big data, machine learning, and deep learning. With over 24 years of expertise in software engineering, IT management, system architectures, and data processing, he also conducts research in big data and machine learning applications in public finance. Nabil has published scientific articles and actively participates in conferences. He can be contacted at email: nabil.ourdani@etu.uae.ac.ma or ourdani@gmail.com.

Mohamed Chrayah is a Computer Science Professor at Abdelmalek Essaadi University in Tetuan, Morocco, where he has been serving since 2014. His primary research interests encompass machine learning, big data, and E-learning, with a particular focus on social media analysis. He earned his Ph.D. in computer science from the same university in 2013 and holds an engineering degree from the National Institute of Statistics and Applied Economics in Rabat, Morocco, achieved in 2006. Mohamed has authored numerous publications in international journals and actively contributed to organizing workshops at prestigious international conferences. He also worked as a lecturer, research fellow, senior developer, and system administrator. He can be contacted at email: m.chrayah@uae.ac.ma.

Noura Aknin is a professor at Abdelmalek Essaadi University since 2000, co-founded IEEE Morocco Section in 2004. She's a member of IEEE's Communications and Computer Societies. As head of the Research Unit for Information Technology and Modeling Systems, she coordinates the Bachelor of Network Administration and Security of Information Systems program. She's actively involved in various research projects, including European ones like Tempus (RIFAINES, Master Computer Engineering, Master ISIS, EOLES), and the Tuniso-Moroccan project on Cloud Computing. Pr. Noura AKNIN is also responsible for several AECI-funded projects and manages the Mobile Networks project funded by "Maroc Telecom." She participates in organizing and scientific committees for multiple symposia and conferences. Her research focuses on telecommunications, mobile and wireless communications, and Web 2 applications. She can be contacted at email: noura.aknin@uae.ac.ma.